Virology 488 (2016) 169–178

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Innate immune system activation by viral RNA: How to predict it?

M. Kondili [a,b], M. Roux [b], N. Vabret [c], M. Bailly-Bechet [a,b,*]

[a] Atelier de Bioinformatique, Université Pierre et Marie Curie - Paris VI, 4 place Jussieu, 75005 Paris, France
[b] Laboratoire Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, CNRS, UMR5558, Bâtiment Gregor Mendel, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne Cedex, France
[c] Unité de Génomique Virale et Vaccination, CNRS UMR-3569, Institut Pasteur, Paris, France

ARTICLE INFO

Article history:
Received 10 June 2015
Returned to author for revisions
24 October 2015
Accepted 10 November 2015

Keywords:
RNA virus
Innate immune system
Activation
Meta-analysis

ABSTRACT

The immune system is able to identify foreign pathogens via different pathways. In the case of viral infection, recognition of the viral RNA is a crucial step, and many efforts have been made to understand which features of viral RNA are detected by the immune system. The biased viral RNA composition, measured as host–virus nucleotidic divergence, or CpG enrichment, has been proposed as salient signal. Peculiar structural features of these RNA could also be related to the immune system activation. Here, we gather multiple datasets and proceed to a meta-analysis to uncover the best predictors of immune system activation by viral RNA. "A" nucleotide content and Minimum Folding Energy are good predictors, and are more easily generalized than more complex indicators suggested previously. As RNA composition and structure are highly correlated, we suggest further experiments on synthetic sequences to identify the viral RNA sensing mechanisms by immune system receptors.

Introduction

Part of the efficacy of the immune system relies on its ability to specifically detect the presence of foreign pathogens during the early stages of infection. In mammals, this function is performed by a subset of cellular receptors called Pattern Recognition Receptors (PRRs). These PRRs have the ability to sense microbial structures absent from the host, classically defined as Pathogen-Associated Molecular Patterns (PAMP, Iwasaki, 2012). In the case of infection by RNA viruses, RNA may be the first and only microbial PAMP produced throughout the larger part of the replication cycle. Thus, a thorough detection of foreign RNA upon viral entry is critical for eliciting an efficient antiviral response.

In the past two decades, an intensive research effort has been deployed to identify viral RNA features that were specifically detected by the innate immune system. This led to the characterization of several RNA PAMP and the receptors they activate. This is the case of RNA molecules bearing a 5′−triphosphate group, a structure found in the genomic RNA or in replication intermediates of many viral families. This 5′−PPP moiety can activate various PRR such as RIG-I, PKR or IFIT-1 and IFIT-5 (Hornung et al., 2006; Pichlmair et al., 2006; Rao Nallagatla, 2007; Pichlmair et al., 2011). Additionally, the absence of methylation residues on foreign RNAs was also shown to activate other PRR such as MDA5 (Züst et al., 2011) or IFITs (Daffis et al., 2010). Finally, RNA secondary folding that form double stranded RNA molecule can activate different receptors such as MDA5, PKR, and RNA helicases of the DEAD box family (reviewed in Vabret et al., 2014).

Several recent studies have demonstrated a link between the activation of innate immune system and properties intrinsic to viral RNA sequences. Following the early work from S. Karlin on dinucleotide abundances and genomic signatures (Karlin and Burge, 1995; Karlin, 1998; Karlin et al., 1994), Greenbaum et al. first demonstrated that in influenza genomes, CpG motifs in an A/U-rich RNA have immunostimulatory properties (Greenbaum et al., 2009; Jimenez-Baranda et al., 2011). This sequence motif is underrepresented in both ssRNA viruses and host innate immune gene mRNA, and its frequency in influenza virus genomes decreased over decades of human adaptation and reflects viral transitions from avian to human hosts (Greenbaum et al., 2008). This motif was shown to induce type I interferon (IFN-I) secretion in a specific subset of immune cells through recognition by the innate sensor Toll-like receptor 7 (Jimenez-Baranda et al., 2011). This work has recently led to a statistical physics modelling of dinucleotide content evolution in RNA viral genomes as an equilibrium between entropic and selective forces (Greenbaum et al., 2014). The group of P. Simmonds also studied the impact of modifying CpG and UpA dinucleotide frequencies in viral genomes. They produced mutants of the picornavirus echovirus 7, by increasing CpG and UpA genome frequencies, and showed that these mutants displayed impaired replication kinetics compared to the wild-type (wt) virus. The use of kinase inhibitors suggested that this

* Corresponding author.
E-mail address: marc.bailly-bechet@univ-lyon1.fr (M. Bailly-Bechet).

http://dx.doi.org/10.1016/j.virol.2015.11.007
0042-6822/© 2015 Elsevier Inc. All rights reserved.

inhibition occurred through a yet undescribed pattern recognition receptors that respond to RNA composition (Atkinson et al., 2014). The same group also analyzed the Minimum Folding Energy Difference (MFED) of RNA viral sequences, which is defined as the excess or lack of folding of the viral RNA sequence when compared to the expectation due to its di-nucleotidic composition (Davis et al., 2008). They identified large secondary structures called GORS (Globally Ordered RNA Structure), observed in many RNA virus families. They measured an inverse correlation between the presence of these GORS and the ability of an RNA sequence to activate the innate immune system (Witteveldt et al., 2014). Our group focused on nucleotide composition of viral sequences. Most RNA virus families show important sequence bias compared to human transcriptome, including in their nucleotide ratio (van der Kuyl and Berkhout, 2012), dinucleotide frequency (Greenbaum et al., 2008) and codon usage bias (Jenkins and Holmes, 2003; Belalov and Lukashev, 2013). These sequence biases are specific to each viral family and nucleotide composition analysis can be used to determine the origin of viral sequence inside the same viral order (Kapoor et al., 2010). Therefore, we hypothesized that specific properties associated to viral nucleotide composition could be recognized by dedicated PRR as patterns absents from host RNAs. Following this rationale, we have shown that RNA sequences derived from lentiviral genomes induced an IFN-I response, and the extent of this response correlated to the nucleotide divergence of the RNA sequence, defined by the chi-square distance between host and virus nucleotidic frequencies (Vabret et al., 2012). We also demonstrated that a synthetic virus with a reduced nucleotide divergence and the same replicative capacity is less immunostimulatory (Vabret et al., 2014).

Altogether, these approaches have put forward relations between sequence-specific RNA characteristics and immune system activation. The immune system may not be able to distinguish all of these characteristics at the molecular level, but only able to detect some of them, the others being spuriously found in experiments because they are correlated to the meaningful ones. In this report we gather different datasets where both viral RNA sequences and interferon activation measurements were available. We then perform a meta-analysis in order to characterize the compositional or structural RNA features that can best be used as predictors of the host innate immune system activation.

Results

Meta-analysis on wt viral sequences

To uncover the more reliable predictor of interferon activation among those which have been put forward by previous experiments, we performed a joint analysis of available experimental data. As experimental setups may vary and experimental measurements cannot be directly compared, we chose to work with experiments spanning at least 10 viral sequences, to be able to get information from the relative activation values in each experiment. We then worked with a set of viral sequences coming from 3 publications: two from Vabret et al. (2012, 2014), and one from P. Simmonds group (Witteveldt et al., 2014). These publications respectively study interferon production in response to stimulation by HIV-1 or SIVmac239 sequences (“HIV 2012” and “SIV 2014” datasets), and two wide-spectrum sets of RNA viruses (“Vabret 2014” and “Witteveldt 2014” virus datasets). The “Full” dataset combines them for a total of 107 wt viral sequences with experimentally measured interferon activation values, that were normalized between datasets before comparison. For each sequence we computed various compositional predictors – base content, nucleotidic and dinucleotidic divergence, dinucleotidic enrichment ratio for CpG and UpA dinucleotides –, and structural predictors: MFE, MFED, as well as the percentage of unpaired bases in the folded RNA. To be comparable, MFE and MFED were averaged on 300 bp subsequences, and should then be considered as per nucleotide measurements, independent of the considered RNA length.

We first investigated the performance of each predictor on each dataset separately. Results are shown in Fig. 1 and Table 1. There is a wide range of variability in activation values (y-axis of Fig. 1), except in the “Witteveldt 2014” dataset, where we observed a threshold between sequences showing a low activation level and four highly activating ones: Bunyamwera, Sindbis, Sendai and Hepatitis A viruses. Then, any predictor should reflect this threshold in order to strongly correlate with activation values; note however that using Spearman correlations we obtain the same qualitative and quantitative results (e.g. Supp. Table S1). Table 1 indicates that data coming from “Witteveldt 2014” and “Vabret 2014” datasets show higher correlations with structural indicators and A content, while “SIV 2014” data systematically gets a lower correlation level with interferon production. “HIV 2012” data offer no such simple description, with nucleotidic divergence, dinucleotidic divergence and MFE being the best predictors, but MFED not being strongly correlated to activation. Note that for the HIV and SIV datasets, experimental SHAPE values measuring local levels of RNA folding are available (Watts et al., 2009; Pollom et al., 2013); we tested for the correlation between these SHAPE values and interferon activation (Table 1, last line), but resulting correlations are non-significant. In conclusion, different datasets call for different best predictors, emphasizing the need for a more global analysis of these experiments.

To gather power in our study we hypothesized that, if the same factors are activators of the immune system in the 4 datasets, we could test those factors on the “Full” dataset (Table 1). This approach led to MFE as the best structural predictor. As for compositional indicators, A content, nucleotidic divergence and dinucleotidic divergence are relatively close, with a R correlation value slightly smaller than for MFE. We see that the correlations observed on the “Full” dataset are smaller than those observed on individual datasets: this indicates that the normalization procedure we used to study at once multiple datasets may not be sufficient to make them perfectly comparable, and that the real average activation levels (or variability in them) may be different between datasets.

Lentiviruses genome have a highly biased nucleotide composition, with up to 40% of their sequence composed of A-nucleotide (Berkhout et al., 2002; van der Kuyl and Berkhout, 2012). This bias has been explained by dNTP pool imbalance during reverse transcription and by the antiviral activity of the cellular Apobec 3G (A3G) cytidine deaminase, which mutates G to A in HIV provirus (Deforche et al., 2007). Concerned that this bias, combined to the number of lentiviruses sequences, could alter the conclusion of our results in the Full dataset, we also tested all predictors on a reduced dataset composed only of Wittelveldt 2014 and Vabret 2014 data (Supp. Table S2). This did not change our overall conclusions but improved the prediction ability of CpG, highlighting its potential importance, particularly in unbiased genomes.

To understand how the different predictors correlate with each other in the “Full” dataset, we proceeded to a principal component analysis on them (Fig. 2). This technique projects a set of variables on a two-dimensional space, such that predictors being correlated positively (resp. negatively) appear as arrows sharing a common direction, and being in the same (resp. opposite) sense. Uncorrelated predictors' arrows are orthogonal. The projection of these data on two dimensions conserves more than 70% of the global variability, a value high enough to draw meaningful conclusions. Most predictors correlate with the first axis of the projection. G and U content do not, which is interesting, as those parameters

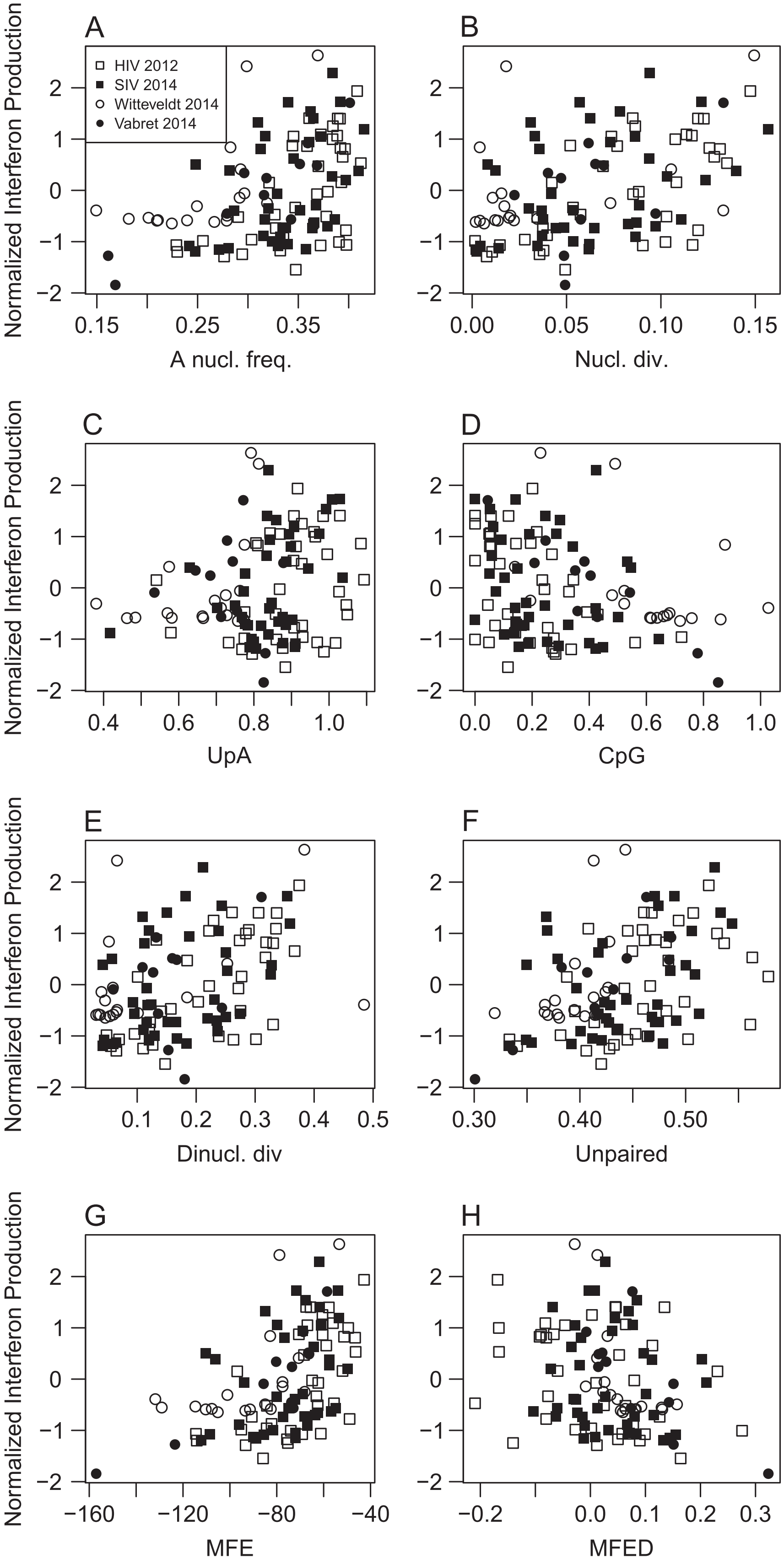

Fig. 1. Relation between compositional or structural predictors and interferon activation in the "Full" dataset. *x*-axis is the predictor, *y*-axis is the normalized interferon production. Point type indicates data origin as shown in box A.

Table 1
Pearson correlation coefficients R and corresponding p values of association between predictors and interferon activation. Cases with a p value smaller than 0.01 are in bold.

Predictors	Datasets				
	HIV 2012	SIV 2014	Witteveldt 2014	Vabret 2014	Full
Nucl. Div.	$\mathbf{R=0.644}$ $\mathbf{p<1.3\times10^{-5}}$	R=0.359 $p<2.3\times10^{-2}$	R=0.423 $p<0.08$	R=0.489 $p<0.13$	$\mathbf{R=0.456}$ $\mathbf{p<7.9\times10^{-7}}$
A	$\mathbf{R=0.577}$ $\mathbf{p<1.6\times10^{-4}}$	R=0.368 $p<2\times10^{-2}$	$\mathbf{R=0.631}$ $\mathbf{p<5\times10^{-3}}$	$\mathbf{R=0.894}$ $\mathbf{p<2.1\times10^{-4}}$	$\mathbf{R=0.462}$ $\mathbf{p<5.4\times10^{-7}}$
CpG	$\mathbf{R=-0.432}$ $\mathbf{p<6.8\times10^{-3}}$	$R=-0.178$ $p<0.28$	$R=-0.427$ $p<0.078$	$\mathbf{R=-0.929}$ $\mathbf{p<3.5\times10^{-5}}$	$\mathbf{R=-0.310}$ $\mathbf{p<1.2\times10^{-3}}$
UpA	R=0.229 $p<0.17$	$\mathbf{R=0.404}$ $\mathbf{p<9.7\times10^{-3}}$	R=0.447 $p<0.063$	$R=-0.188$ $p<0.58$	R=0.238 $p<1.4\times10^{-2}$
Din. Div.	$\mathbf{R=0.651}$ $\mathbf{p<9.5\times10^{-6}}$	R=0.374 $p<1.8\times10^{-2}$	R=0.352 $p<0.15$	R=0.247 $p<0.47$	$\mathbf{R=0.425}$ $\mathbf{p<5\times10^{-6}}$
MFE	$\mathbf{R=0.615}$ $\mathbf{p<4\times10^{-5}}$	R=0.327 $p<4\times10^{-2}$	R=0.588 $p<1.1\times10^{-2}$	$\mathbf{R=0.856}$ $\mathbf{p<7.7\times10^{-4}}$	$\mathbf{R=0.497}$ $\mathbf{p<5.3\times10^{-8}}$
MFED	$R=-0.306$ $p<0.062$	$R=-0.097$ $p<0.56$	$\mathbf{R=-0.625}$ $\mathbf{p<5.6\times10^{-3}}$	$\mathbf{R=-0.774}$ $\mathbf{p<5.2\times10^{-3}}$	$\mathbf{R=-0.296}$ $\mathbf{p<2\times10^{-3}}$
% Unpaired	$\mathbf{R=0.441}$ $\mathbf{p<5.6\times10^{-3}}$	R=0.383 $p<1.5\times10^{-2}$	R=0.516 $p<2.9\times10^{-2}$	$\mathbf{R=0.848}$ $\mathbf{p<9.8\times10^{-4}}$	$\mathbf{R=0.421}$ $\mathbf{p<6.2\times10^{-6}}$
SHAPE	$R=-0.320$ $p<0.061$	R=0.258 $p<0.14$	NA NA	NA NA	NA NA

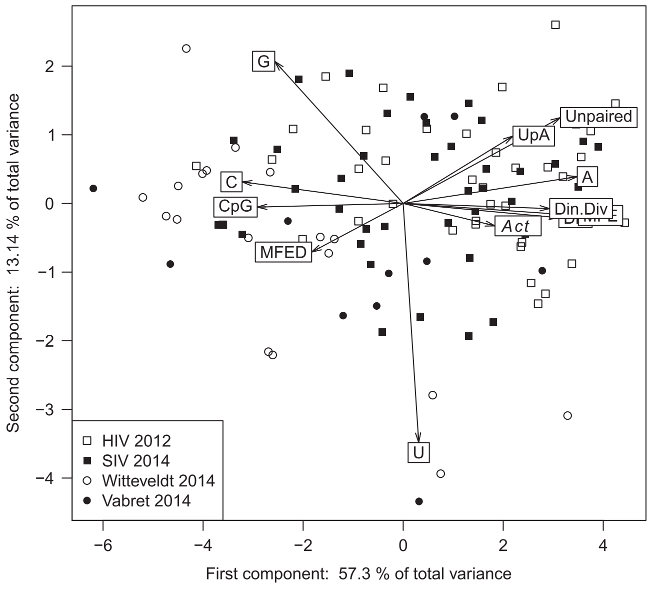

Fig. 2. Principal component analysis of the "Full" dataset. All predictors' variables were used for the analysis. Normalized interferon activation (*Act*) was projected on this space afterwards and is indicated in italic. Point type indicates data origin. The three overlapping arrow labels on the right are dinucleotidic divergence, nucleotidic divergence and MFE.

were included in our analysis for completeness but have not been suspected to be related to the immune system activation. Indeed, projection of the interferon activation data on this plane, as a supplementary variable, shows that this activation mainly correlates with the first axis. As expected, A content, nucleotidic divergence, dinucleotidic divergence and MFE correlate positively with activation, while MFED and CpG enrichment ratio correlate negatively with interferon production.

Looking at the same data in a more quantitative way, we saw many significant correlations between those predictors (Table 2). The highest level of correlation (R=0.97) is observed between nucleotidic and dinucleotidic divergence, indicating that little information is gained on the "Full" dataset by looking at nucleic frequencies of higher order. The second higher correlation (R=0.92) is observed between A content and MFE, two of the best predictors uncovered in the previous analyses. Moreover, many other couples do correlate above the R=0.7 level, showing that there may well be a confusion about the biological causes of interferon activation: an experiment could detect a predictor as significantly related to the immune system activation while it would only be of proxy for another one. For example, all predictors except for U content have an absolute correlation coefficient higher than 0.47 with MFE, making this predictor a good "summary candidate" for all of them.

Activation changes in wt/synthetic sequence pairs

A content and MFE are good predictors of the interferon activation in wt sequences, but are very highly correlated. This correlation makes sense as an A-rich RNA will make more A–U pairs while folding, lowering its MFE; moreover, we can see in Table 2 that higher A content also results in less Watson-Crick base pairs in the folded RNA. This global property makes it difficult to distinguish between A content and MFE on the basis of our previous analyses. We then decided to analyze three other datasets, coming from the same papers, but made of close pairs of sequences, one wt and the other a synthetic sequence derived from the wt. By comparing the impact of relatively small sequence modifications that affect either the A content, or the MFE, we wanted to get a more precise insight on the biological impact of sequence composition and structure on interferon production.

We first re-analyzed data from Fig. 5D of Witteveldt et al. (2014), which is a set of three wt/synthetic pairs of sequences from the MNV and HCV viruses. The methodology used to create synthetic sequences kept the nucleotidic and dinucleotidic composition constant. Then, differences in interferon production can only be attributed to differences in structural characteristics, MFE and MFED. For the two first pairs – noted US for unstructured in the original paper, as these sequences have a lower MFED than the wt – the activation is 3–5 times stronger in the synthetic sequences, while MFE is stronger and MFED is reduced (Table 3). These results point towards an effective role of RNA structural features in immune system activation. The last synthetic sequence, MNV-RS, has been designed to have the same

Table 2
Pearson correlation coefficients R between all predictors used in this study, on the "Full" dataset. Lines separate compositional and structural predictors. The highest correlation between compositional and structural indicators is shown in bold.

Predictors	A	C	G	U	Div	CpG	UpA	Din. Div	MFE	MFED	%Unpaired
A	1.00										
C	−0.88	1.00									
G	−0.64	0.51	1.00								
U	−0.08	−0.21	−0.49	1.00							
Div	0.71	−0.62	−0.59	0.08	1.00						
CpG	−0.79	0.80	0.46	−0.04	−0.57	1.00					
UpA	0.50	−0.49	−0.23	−0.08	0.44	−0.43	1.00				
Din. Div	0.63	−0.54	−0.52	0.06	0.97	−0.51	0.52	1.00			
MFE	**0.92**	−0.85	−0.73	0.13	0.68	−0.70	0.47	0.61	1.00		
MFED	−0.35	0.38	0.18	−0.02	−0.28	0.27	−0.29	−0.29	−0.60	1.00	
% Unpaired	0.85	−0.65	−0.49	−0.27	0.67	−0.57	0.46	0.61	0.85	−0.50	1.00

Table 3
Structural values MFE and MFED of the 5 HCV and MNV sequences from Fig. 5D of Witteveldt et al. (2014). wt stands for wild-type and US for unstructured, i.e. sequences with the same nucleotidic, di-nucleotidic and coding content than wt, but with randomly permuted nucleotides. Note that indeed MFED of the US sequences is close to 0, which is expected from this design procedure. RS sequence stands for re-stabilized, indicating that the sequence was selected after a specific nucleotide permutation keeping a MFED value close to the native one. Note that indeed MFED of RS sequence is close to MFED of the wt.

Predictors	Sequences				
	HCVwt	HCV-US	MNVwt	MNV-US	MNV-RS
MFE	−121.90	−112.07	−114.72	−108.00	−114.37
MFED	0.068	−0.0146	0.096	0.036	0.10

Table 4
Pearson correlation coefficients R and corresponding p values of association between the differences in predictors inside each sequence pair and the relative differences in interferon activation, in the "Viral Pairs" dataset.

Stat.	Predictors				
	ΔA nucl.	Δ Nucl. Div.	Δ Din. Div.	ΔMFE	ΔMFED
R	0.831	0.135	0.018	0.788	−0.662
p	$<2.9\times10^{-3}$	<0.71	<0.96	$<6.9\times10^{-3}$	$<3.8\times10^{-2}$

MFED as the wt sequence. It also has the same MFE, and nonetheless show a very strong increase in interferon production. Then, modifications of the wt sequence with no measurable influence on nucleotidic content nor on structural indicators have a significant effect on immune system activation. These results, which mirror those primarily obtained by Witteveldt et al. (2014), could be a consequence of experimental inaccuracy; apart from that possibility, they indicate that there are cases where MFE or A content may not be used to predict interferon production. It is a possibility that the design of the MNV-RS sequence has modified other characteristics of the viral RNA, that are not directly related to nucleic content or folding properties.

Another dataset we could study is composed of 10 viral sequences pairs, with one wt sequence and one codon-usage optimized sequence, in multiple RNA viruses from Vabret et al. (2014). We measured differences in interferon activation between both sequences inside each pair, and compared it to difference in the predictors. Corresponding results are shown in Table 4 and Fig. 3. ΔMFE is a better predictor than ΔMFED, and ΔA is in between; moreover, graphs B and C are very similar in Fig. 3, showing that the correlation between A content and MFE seem to apply on these data.

However, here, an increase in both nucleotidic and dinucleotidic divergence does not significantly correlate with the increase in interferon production. This is intriguing, as a decrease of nucleotidic divergence of SIVmac239 results in a synthetic sequence (SIVopt1) which has been experimentally shown to be less activator of the innate immune system, in biologically realistic conditions (Vabret et al., 2014, Fig. 5B). Then, we decided to study in more detail the couple of sequences SIVmac239/SIVopt1. These sequences differ at 73 positions in the segment 3645–5001, inside the *pol* sequence. Characteristics of both wt and synthetic sequences are summed up in Table 5. Again, the decrease in interferon activation in the synthetic sequence is concordant with a decrease both in A nucleotidic frequency, and in MFE. However, here the MFED decreases with decreasing activation, which is not in agreement with the previous predictions and the experimental results from Witteveldt et al. (2014).

We could proceed further on this couple of sequences, as the RNA structure of the SIVmac239 genome is known (Pollom et al., 2013). This allows for a numerical study of the structural consequences of the changes between wt and synthetic sequences, in a more precise way than estimating MFE and MFED. By folding separately and incrementally the known folding domains of SIVmac239, we were able to predict the most probable structure of SIVopt1. Then, we compared the respective foldings of SIVopt1 and SIVmac239. The main result is that the full-length MFE of SIVopt1 is slightly higher (−387.5 vs −390, less than 1% increase) than the MFE of SIVmac239, and both structures (Figs. S1 and S2) are qualitatively similar. Note that these values are not numerically comparable with our previous MFE values based on averages of 300 bp segments, but that their computation requires the knowledge of the global RNA fold. So the 73 modifications in the SIVmac239 *pol* sequence have almost no consequence on the RNA structure of this subpart of the SIVmac239 RNA, despite the differences observed when using the standard way of measuring sequences MFE and MFED. However, experiments have shown a significant decrease of interferon production in SIVopt1, pointing towards a sensing mechanism based here on nucleotide composition. In particular the A content has been reduced by 4.2% in this synthetic sequence, a relative diminution of more than 10% relative to the wt.

All the predictors we studied until now are large-scale RNA sequence characteristics. However, short and local PAMP may activate the immune system. The three last datasets, composed of close sequence pairs with different levels of immune system activation, could be used to investigate the presence of short immune activator patterns on the RNA sequence. This approach could detect, e.g., CpG in an A/U context as described by Greenbaum et al. (2009). If such patterns are recognized by the innate immune system, they should be among those disrupted or formed by substitutions inside each sequence pair, and then should be detected by a statistical approach. These patterns, unless necessary for host infection, should progressively disappear through evolution; but we nonetheless expect that in pairs of close sequences, those containing more activator patterns would be experimentally characterized by a higher immune system activation.

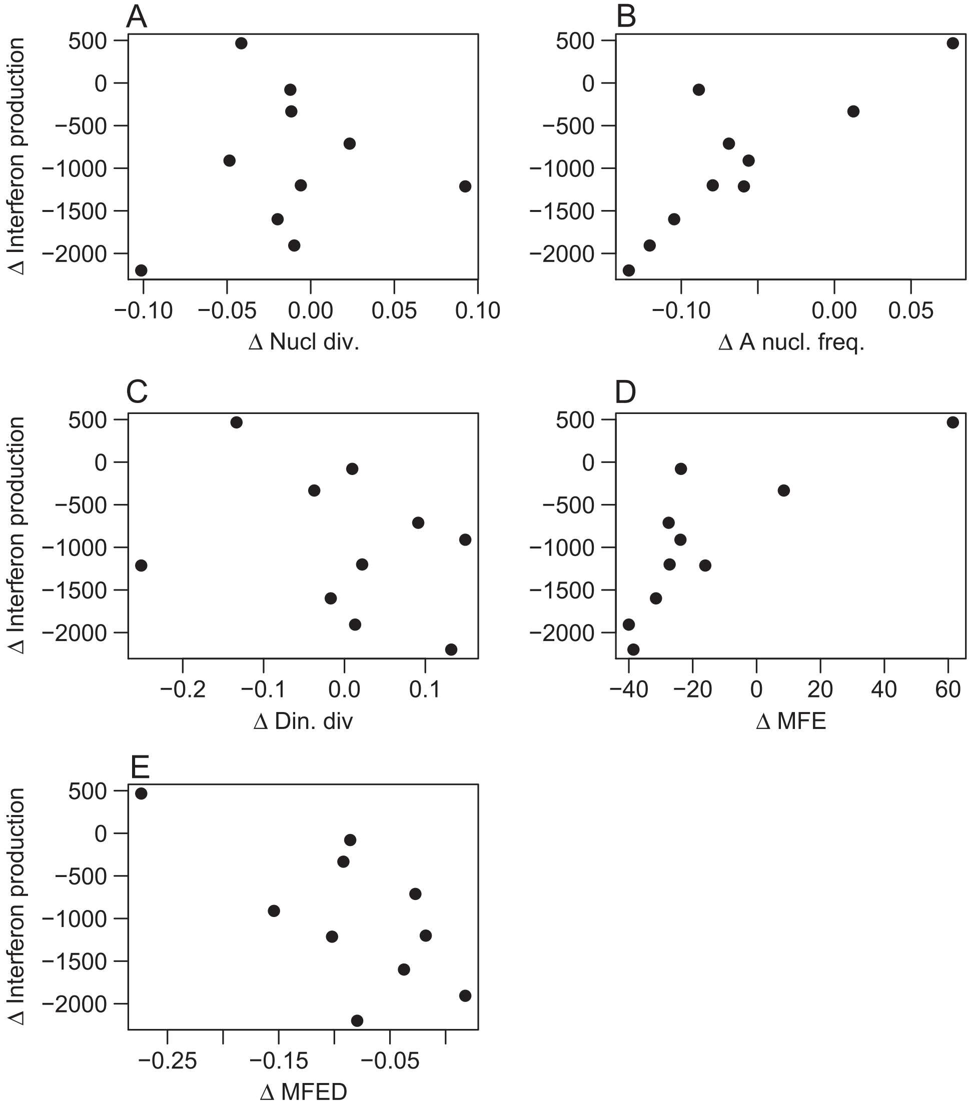

Fig. 3. Relation between the intra-pairs differences in the 5 genomic indicators and differences in interferon production, in the "Viral Pairs" dataset. Genomic indicators are on the *x*-axis, interferon production is on the *y*-axis. Interferon differences were not normalized here as all data come from the same experiment. Note that the peculiar point with a positive *Δ* Interferon production is HCVcore, a viral sequence with a very high wt GC content. See Vabret et al. (2014) for more precisions.

Table 5
Compositional and structural predictors value for SIVmac239 and SIVopt1 *pol* segment.

Sequences	Predictors				
	A nucl.	Nucl. Div.	Din. Div.	MFE	MFED
SIVmac239, pos 3645 to 5001	0.391	0.113	0.289	−56.1	0.00137
SIVopt1, pos 3645 to 5001	0.349	0.051	0.141	−65.5	−0.0169

We selected short patterns, more frequent in the more activating sequence of the pair, and clustered them to get consensus patterns. This allowed us to define 5 consensus patterns: AAAGA, ACAUU, AAAUG, CAAUA and UGGGAA. Then, we counted all occurrences of those consensus patterns, allowing for one error, in all sequence pairs. Results are shown in Fig. 4. There is no single consensus pattern systematically more frequent in the more activating sequence. However, the cases where the consensus activator pattern is less present in the more activating sequence (e.g. ACAUU in HCV E1E2 pair, or AAAGA in MNV-RS/wt) always show small occurrence numbers for the pattern, both for the wt and modified sequence. These small numbers indicate that any methodological change in counting the pattern occurrences, such as allowing for less errors or changing the consensus threshold, may affect the results.

Then, it cannot be excluded that one of those patterns is indeed detected by the immune system, and that experimental or methodological inaccuracies prevent uncovering it more precisely. However, all those patterns are A/U rich (minimum 50%) but do not contain any CpG motif, preventing us to relate our results to those of Greenbaum et al. (2009).

Discussion

Several intrinsic sequence properties have been shown to be responsible for innate immune activation by RNA molecules. This is the case of nucleotide ratio, dinucleotide frequencies or global viral RNA structures (Atkinson et al., 2014; Vabret et al., 2012; Jimenez-Baranda et al., 2011; Witteveldt et al., 2014). In this study, we benefited from similar experimental methods performed by independent investigators to compare these properties in their capacity to correlate

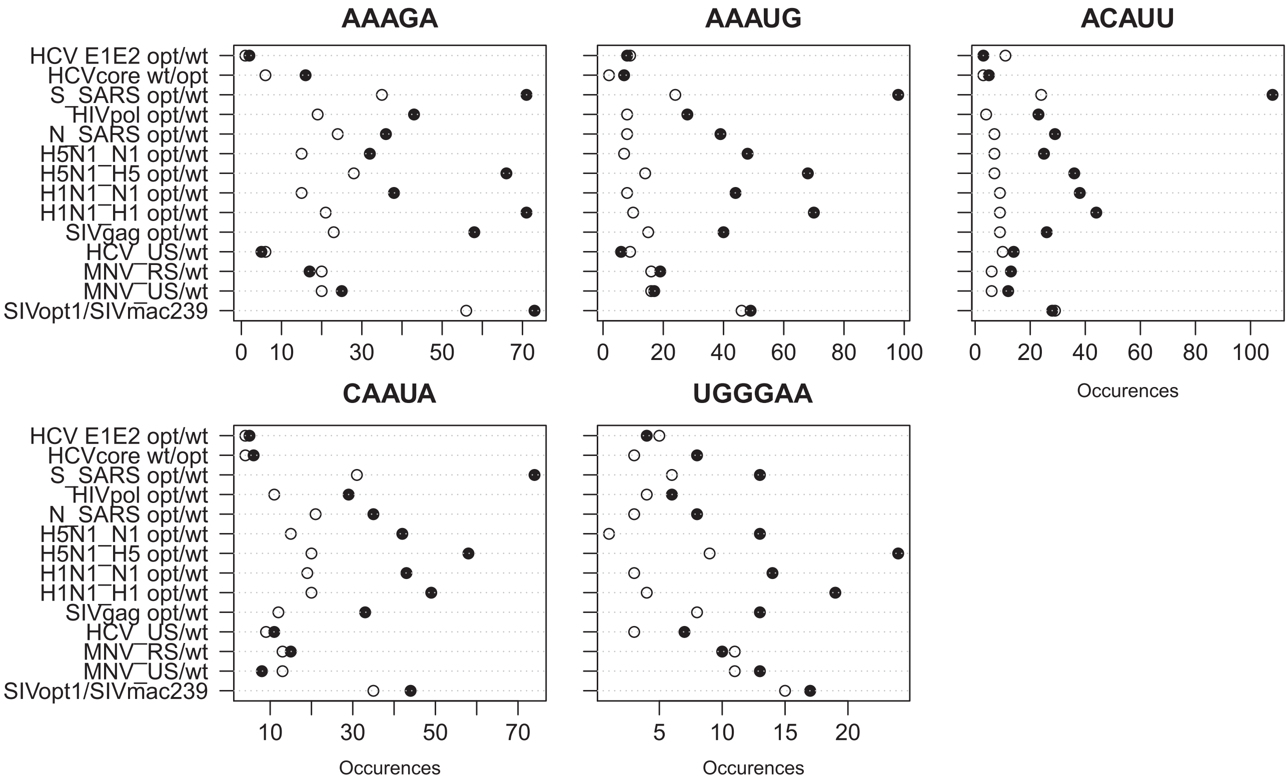

Fig. 4. Occurrences of the five patterns found as putatively linked to interferon activation in the combined "Viral Pairs", SIVmac239/SIVopt1 and MNV/HCV pairs datasets. Occurrence number is on the *x*-axis, and sequence names are indicated on the *y*-axis: name before the "/" symbol is the less activating sequence of the pair, name after is the more activating one. On the graph, black points stand for the more activating sequence, white ones for the less activating one. All occurrence counts have been computed with a 1 error tolerance in the pattern.

with innate immune system activation. Among the predictors based on sequence composition, A content is the best correlate and should then be used instead of the nucleotidic divergence proposed in Vabret et al. (2012). Concerning structural predictors, MFE is consistently more accurate than MFED. Moreover, it is easier and faster to compute, and we recommend using it instead of MFED.

A complete understanding of the cellular mechanisms that govern differential recognition of microbial RNA in response to its sequences is still missing. In this context, determining whether structural or compositional determinants better predicts the innate immune activation is critical in order to define the mechanisms involved. By nature, RNA secondary structure and nucleotide composition are highly entangled. In the datasets used in this study, MFE and A-content are very highly correlated (Table 2). Analysis of the MNV-RS sequence from Witteveldt et al. (2014) indicates that immune system level of activation can vary even when modifications of the wt sequence keep both MFE and nucleotidic content constant. Other results from the same work show that an increase in MFE, which means less structured RNAs, can increase the immune response. However comparison of the folding structures of SIVopt1 and SIVmac239 shows that with a quasi-constant MFE and a qualitatively identical fold, a decrease in A content is linked to a decrease in interferon production. Therefore, two sensing mechanisms, based on RNA composition and structure, may be necessary to explain these data.

Recently, a paper devoted to specific molecular mechanisms of viral RNA recognition completed our understanding of compositional vs structural predictors. The work by Chiang et al. (2015) designed two synthetic RNA containing 5′−PPP moieties (labeled M5 and M8) that have enhanced abilities to induce IFN-β through the RIG-I pathway. We computed both A content and MFE of these sequences relative to the WT sequence they were derived from. M5 has a slightly higher A content than the WT RNA (30.8% vs 29.2%), and a lower MFE per nucleotide, i.e. is more stable. M8 had a much higher A content (37.2%) and is also more stable than the WT sequence. The authors hypothesized that the structural stability of the M8 sequence induced a higher half-life of the RNA and that sequence modifications improved the ability of RNA to bind to RIG-I and to induce its oligomerization. Yet, they do not discuss the higher A content as a possible source of immunostimulation.

The fact that higher structural stability can also increase RNA half-life in the cell (Kohlway et al., 2013) emphasizes the importance of carefully designing experimental systems in order to discriminate the innate immune stimulation by RNA from its other functions in the cell. Specific RNA sequences can modulate its interaction with the host translation machinery or the RNA quality control and degradation pathways (Abernathy and Glaunsinger, 2015). This has been suggested to explain the difference in folding free energy between avian and human influenza virus RNAs (Brower-Sinning et al., 2009). Viral RNA sequences can also provide a molecular signature that is recognized during virus replication (van der Kuyl and Berkhout, 2012), for example to monitor RNA packaging (Sebla et al., 2015). Finally, in relevant species, specific sequence modification can result from the viral escape to RNA interference and cellular intrinsic immunity (Kemp and Imler, 2009).

Our understanding of how the immune system distinguishes between viral and self-RNAs will continue to improve over time. Designing robust experimental systems that allow us to discriminate the role of each RNA feature on innate immune activation remains one of its challenge.

Material and methods

Data sources

Datasets where both viral RNA sequences and interferon activation values for at least 10 RNA viruses were available were selected for the study. The "HIV 2012" dataset is composed of 38 RNAs 500 bp long, overlapping and spanning the complete HIV-1 hxb2 clone genome sequence (accession number K03455), from Vabret et al. (2012). The "SIV 2014" dataset is composed of 40 RNAs 500 bp long, overlapping and spanning the complete SIVmac239 clone genome sequence (accession number M33262), from Vabret et al. (2014). The "Witteveldt 2014" dataset contains all 18 RNA sequences studied in Witteveldt et al. (2014). The "Vabret 2014" dataset contains 11 RNA sequences of wild-type viruses from Fig. 1 of Vabret et al. (2014). Dataset called "Full" is the aggregation of the 4 previously described datasets. In each of those four datasets, interferon activation values were normalized to have 0 mean and 1 variance, to make data comparable.

The "Viral Pairs" dataset is composed of 10 pairs of viral sequences (wt and codon optimized) from Vabret et al. (2014), Fig. 1 (HIV-1 Gag sequences were removed due to the lack of corresponding interferon activation value). Wild-type sequences are identical to those of the "Vabret 2014" dataset. Accession numbers and nucleotidic composition of all sequences in these datasets are given in Supp. File 1. Both wt partial sequences of Murine Norovirus (MNV) and Hepatitis C virus (HCV), and the corresponding synthetic sequences (HCV-US, MNV-US, MNV-RS) from Witteveldt et al. (2014), Fig. 5D, were graciously provided by the authors. HCV sequences are Hepatitis C Virus sequences (Genbank ID AJ238799, pos. 6529 to 7635); MNV are Murine Norovirus 3 sequences from the pt7-MNV3 clone (Genbank ID JQ658375, pos. 1147 to 2467). The optimized, synthetic genome sequence of the SIVmac239 genome called SIVopt1 (accession number KJ152770) is from Vabret et al. (2014). The only part used in the computations (Table 5) is the optimized subpart of the genome, from positions 3645 to 5001 of the wt sequence.

Predictors computation

Nucleotidic divergence was computed according to Vabret et al. (2012), using the following formula:

$$\text{Nucl. Div(host, virus)} = \sum_{b \in \{A,C,G,U\}} \frac{(f_v(b)-f_h(b))^2}{f_h(b)^2},$$

with $f_v(b)$ and $f_h(b)$ being the respective virus and host frequencies of nucleotide b. As all immune system activation experiments were done in human cells, host frequencies were set to the human genome nucleotidic frequencies.

Dinucleotidic divergence, the equivalent of the previous measure for dinucleotide content, was computed using the same rationale:

$$\text{Din. Div(host}, virus) = \sum_{i \in \{A,C,G,U\}} \sum_{j \in \{A,C,G,U\}} \frac{(f_v(ij)-f_h(ij))^2}{f_h(ij)^2},$$

with i and j being the first and second base of the dinucleotide, f_v and f_h respectively standing for viral and human frequencies.

Dinucleotidic enrichment ratio for CpG and UpA dinucleotides was computed as

$$\text{Dinucl}(ij) = \frac{f_v(ij)}{f_v(i)f_v(j)},$$

with the same notations as in previous formula.

Minimum Folding Energy (MFE) of each RNA was computed using RNAfold (Lorenz et al., 2011) with default parameters, as follows: each RNA was cut in 300 bp long subsequences overlapping by 275 bp, the last subsequence being slightly longer (between 300 and 324 bp). The best folding and corresponding MFE was computed for each subsequence. Then, MFE was computed as the average of the MFE of all the RNA subsequences, the last subsequence MFE being normalized by the corresponding length factor. Tests were made with cuts shorter than 300 bp with no influence on our results.

The typical percentage of unpaired nucleotides in each folded RNA was computed by averaging values measured on the optimal folds predicted by RNAfold, on the set of RNA subsequences 300 bp long and overlapping by 275 bp, for each sequence under study.

Minimum Folding Energy Difference (MFED) was computed in a similar way to Witteveldt et al. (2014). Each RNA was decomposed in subsequences 300 pb long and overlapping by 275 bp, as previously; for each subsequence, 100 random sequences of the same length and same dinucleotidic composition were generated with uShuffle (Jiang et al., 2008), and their MFE was computed with RNAfold. Then, MFED of the original sequence was computed as

$$MFED = \frac{MFE - \langle MFE \rangle_{rand}}{MFE},$$

with *MFE* designing the MFE of the real sequence and $\langle MFE \rangle_{rand}$ the average MFE of the 100 corresponding random sequences. MFE being negative, a positive MFED indicates an excess of structure in the real sequence relative to the randomized ones, and vice versa.

For the "SIV 2014" and "HIV 2012" datasets, SHAPE experimental probabilities of interaction per nucleotide are available (Watts et al., 2009; Pollom et al., 2013). Sequences from these experiments were aligned respectively with the HIV hxb2 and SIVmac239 genomes used in Vabret et al. (2012, 2014), and the average value of the SHAPE parameter was computed for each of our 500 bp RNA sequences who was aligned on its full length.

All correlations between predictors and interferon production were computed as Pearson correlations. Principal components analysis (centered and normalized) was performed on the "Full" dataset, using as variables all 4 nucleotide frequencies, UpA and CpG enrichment ratio, nucleotidic and dinucleotidic divergences, MFE and MFED, and the percentage of unpaired nucleotides in the folded RNA. Normalized interferon activation was then projected on the resulting space as a supplementary variable. Analyses were made with R (R Core Team, 2014) using package ade4 (Dray and Dufour, 2007).

SIVopt1 and SIVmac239 folding

To fold the SIVopt1 and SIVmac239 sequences, limits of folding domains were downloaded from Dataset 1 of Supp. Mat. of Pollom et al. (2013). To maintain folding domain integrity, the sequences compared were larger than the interval inside which SIVop1 was optimized, and ranged from position 3497 to 5467, in the SIVMM239 sequence reference (accession number M33262), corresponding to positions 2979 to 4949 in Pollom et al. (2013). Folding domains were ranked by increasing size. Each domain was folded separately using RNAfold with default parameters, having as an input constraint the folding obtained for all its nested domains. Finally, the full 1791 bp sequence was folded, with the constraint of all the pre-computed domains foldings. This was done separately on the SIVopt1 and SIVmac239 sequence, with the same domain boundaries.

Short activator patterns

Short activator patterns on the RNA sequence were identified as follows. The 14 available pairs of sequences composed of a wt sequence and a synthetic one (i.e. the 10 pairs from the "Viral Pairs" dataset, the SIVopt1 optimized segment vs SIVmac239 wt segment, and the 3 MNV/HCV pairs from Witteveldt et al. (2014)) were used. In each pair, the number of occurrences of each pattern was computed using RISO (Carvalho et al., 2006), with patterns defined as follows:

- 4 to 5 letter word with no error,
- 6 to 8 letters word with at most one error.

For the 14 pairs, the differences in pattern occurrence between the more activating sequence and the less activating one, as reported in the corresponding publications, was computed. Then, patterns being strictly less frequent in the less activating sequence than in the more activating one in at least 11 pairs out of 14 were selected. These patterns were aligned using ClustalW (Larkin et al., 2007) and clustered using the hclust R package (R Core Team, 2014). Groups of approximately equal number of patterns were formed based on visual analysis of the clustering tree, and for each of them a consensus pattern was built with a threshold of 40% presence at each position. One group with an undefined consensus sequence (with more than one nucleotide below the 40% threshold) was removed from further analysis.

Acknowledgments

MBB is grateful to B. Billoud and M. Boccara for useful discussions about the project, for G. Achaz in co-tutoring M. Kondili internship, and all ABI people for welcoming him during his sabbatical. We also thank P. Simmonds for furnishing us sequences relative to Fig. 5D of Witteveldt et al. (2014), and two anonymous referees for their useful comments about this manuscript. MBB acknowledges funds from LBBE, UMR 5558, in Lyon, for M. Kondili internship.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2015.11.007.

References

Abernathy, Emma, Glaunsinger, Britt, 2015. Emerging roles for RNA degradation in viral replication and antiviral defense. Virology 479–480 (May), 600–608.

Atkinson, Nicky J., Witteveldt, Jeroen, Evans, David J., Simmonds, Peter, 2014. The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. Nucl. Acids Res. 42 (April (7)), 4527–4545.

Belalov, Ilya S., Lukashev, Alexander N., 2013. Causes and implications of codon usage bias in RNA viruses. PLoS ONE 8 (2), e56642.

Berkhout, Ben, Grigoriev, Andrei, Bakker, Margreet, Lukashov, Vladimir V., 2002. Codon and amino acid usage in retroviral genomes is consistent with virus-specific nucleotide pressure. AIDS Res. Hum. Retrovir. 18 (January (2)), 133–141.

Brower-Sinning, Rachel, Carter, Donald M., Crevar, Corey J., Ghedin, Elodie, Ross, Ted M., Benos, Panayiotis V., 2009. The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus. Genome Biol. 10 (2), R18.

Carvalho, Alexandra M., Freitas, Ana T., Oliveira, Arlindo L., Sagot, Marie-France, 2006. An efficient algorithm for the identification of structured motifs in DNA promoter sequences. IEEE/ACM Trans. Comput. Biol. Bioinform. 3 (2), 126–140.

Chiang, Cindy, Beljanski, Vladimir, Yin, Kevin, Olagnier, David, Ben Yebdri, Fethia, Steel, Courtney, Goulet, Marie-Line, DeFilippis, Victor R., Streblow, Daniel N., Haddad, Elias K., Trautmann, Lydie, Ross, Ted, Lin, Rongtuan, Hiscott, John, 2015. Sequence-specific modifications enhance the broad spectrum antiviral response activated by RIG-I agonists. J. Virol. (May)

Daffis, Stephane, Szretter, Kristy J., Schriewer, Jill, Li, Jianqing, Youn, Soonjeon, Errett, John, Lin, Tsai-Yu, Schneller, Stewart, Zust, Roland, Dong, Hongping, Thiel, Volker, Sen, Ganes C., Fensterl, Volker, Klimstra, William B., Pierson, Theodore C., Mark Buller, R., Gale, Michael, Shi, Pei-Yong, Diamond, Michael S., 2010. 2′−O methylation of the viral mRNA cap evades host restriction by IFIT family members. Nature 468 (November (7322)), 452–456.

Davis, Matthew, Sagan, Selena M., Pezacki, John P., Evans, David J., Simmonds, Peter, 2008. Bioinformatic and physical characterizations of genome-scale ordered RNA structure in mammalian RNA viruses. J. Virol. 82 (December (23)), 11824–11836.

Deforche, Koen, Camacho, Ricardo, Van Laethem, Kristel, Shapiro, Beth, Moreau, Yves, Rambaut, Andrew, Vandamme, Anne-Mieke, Lemey, Philippe, 2007. Estimating the relative contribution of dNTP pool imbalance and APOBEC3G/3F editing to HIV evolution in vivo. J. Comput. Biol. 14 (October (8)), 1105–1114.

Dray, S., Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. J. Stat. Softw. 22 (4), 1–20.

Greenbaum, Benjamin D., Levine, Arnold J., Bhanot, Gyan, Rabadan, Raul, 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. 4 (June (6)), e1000079.

Greenbaum, Benjamin D., Rabadan, Raul, Levine, Arnold J., 2009. Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. PLoS ONE 4 (6), e5969.

Greenbaum, Benjamin D., Cocco, Simona, Levine, Arnold J., Monasson, Rémi, 2014. Quantitative theory of entropic forces acting on constrained nucleotide sequences applied to viruses. Proc. Natl. Acad. Sci. U.S.A. 111 (April (13)), 5054–5059.

Hornung, Veit, Ellegast, Jana, Kim, Sarah, Brzózka, Krzysztof, Jung, Andreas, Kato, Hiroki, Poeck, Hendrik, Akira, Shizuo, Conzelmann, Karl-Klaus, Schlee, Martin, Endres, Stefan, Hartmann, Gunther, 2006. 5′−Triphosphate RNA is the ligand for RIG-I. Science 314 (November (5801)), 994–997.

Iwasaki, Akiko, 2012. A virological view of innate immune recognition. Annu. Rev. Microbiol. 66, 177–196.

Jenkins, Gareth M., Holmes, Edward C., 2003. The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res. 92 (March (1)), 1–7.

Jiang, Minghui, Anderson, James, Gillespie, Joel, Mayne, Martin, 2008. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinform. 9, 192.

Jimenez-Baranda, Sonia, Greenbaum, Benjamin, Manches, Olivier, Handler, Jesse, Rabadán, Raúl, Levine, Arnold, Bhardwaj, Nina, 2011. Oligonucleotide motifs that disappear during the evolution of influenza virus in humans increase alpha interferon secretion by plasmacytoid dendritic cells. J. Virol. 85 (April (8)), 3893–3904.

Kapoor, Amit, Simmonds, Peter, Ian Lipkin, W., 2010. Discovery and characterization of mammalian endogenous parvoviruses. J. Virol. 84 (December (24)), 12628–12635.

Karlin, S., 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr. Opin. Microbiol. 1 (October (5)), 598–610.

Karlin, S., Burge, C., 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11 (July (7)), 283–290.

Karlin, S., Doerfler, W., Cardon, L.R., 1994. Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. 68 (May (5)), 2889–2897.

Kemp, Cordula, Imler, Jean-Luc, 2009. Antiviral immunity in drosophila. Curr. Opin. Immunol. 21 (February (1)), 3–9.

Kohlway, Andrew, Luo, Dahai, Rawling, David C., Ding, Steve C., Marie Pyle, Anna, 2013. Defining the functional determinants for RNA surveillance by RIG-I. EMBO Rep. 14 (September (9)), 772–779.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23 (November (21)), 2947–2948.

Lorenz, Ronny, Bernhart, Stephan H., Höner Zu Siederdissen, Christian, Tafer, Hakim, Flamm, Christoph, Stadler, Peter F., Hofacker, Ivo L., 2011. ViennaRNA package 2.0. Algorithms Mol. Biol. 6, 26.

Pichlmair, Andreas, Schulz, Oliver, Ping Tan, Choon, Näslund, Tanja I., Liljeström, Peter, Weber, Friedemann, Reis e Sousa, Caetano, 2006. RIG-I-mediated antiviral responses to single-stranded RNA bearing 5′−phosphates. Science 314 (November (5801)), 997–1001.

Pichlmair, Andreas, Lassnig, Caroline, Eberle, Carol-Ann, Górna, Maria W., Baumann, Christoph L., Burkard, Thomas R., Bürckstümmer, Tilmann, Stefanovic, Adrijana, Krieger, Sigurd, Bennett, Keiryn L., Rülicke, Thomas, Weber, Friedemann, Colinge, Jacques, Müller, Mathias, Superti-Furga, Giulio, 2011. IFIT1 is an antiviral protein that recognizes 5′−triphosphate RNA. Nat. Immunol. 12 (July (7)), 624–630.

Pollom, Elizabeth, Dang, Kristen K., Lake Potter, E., Gorelick, Robert J., Burch, Christina L., Weeks, Kevin M., Swanstrom, Ronald, 2013. Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs. PLoS Pathog. 9 (4), e1003294.

Rao Nallagatla, Subba, Hwang, Jungwook, Toroney, Rebecca, Zheng, Xiaofeng, Cameron, Craig E., Bevilacqua, Philip C., 2007. 5′−triphosphate−dependent activation of PKR by RNAs with short stem-loops. Science 318 (November (5855)), 1455–1458.

R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sebla, Kutluay B., Trinity, Zang, Daniel, Blanco-Melo, Chelsea, Powell, David, Jannain, Manel, Errando, Paul, Bieniasz D., 2015. Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. Cell 159 (September (5)), 1096–1109.

Vabret, Nicolas, Bailly-Bechet, Marc, Najburg, Valérie, Müller-Trutwin, Michaela, Verrier, Bernard, Tangy, Frédéric, 2012. The biased nucleotide composition of HIV-1 triggers type I interferon response and correlates with subtype D increased pathogenicity. PLoS ONE 7 (4), e33502.

Vabret, Nicolas, Bailly-Bechet, Marc, Lepelley, Alice, Najburg, Valérie, Schwartz, Olivier, Verrier, Bernard, Tangy, Frédéric, 2014. Large-scale nucleotide optimization of simian immunodeficiency virus reduces its capacity to stimulate type I interferon in vitro. J. Virol. 88 (April (8)), 4161–4172.

van der Kuyl, Antoinette C., Berkhout, Ben, 2012. The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus. Retrovirology 9, 92.

Watts, Joseph M., Dang, Kristen K., Gorelick, Robert J., Leonard, Christopher W., Bess, Julian W., Swanstrom, Ronald, Burch, Christina L., Weeks, Kevin M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460 (August (7256)), 711–716.

Witteveldt, Jeroen, Blundell, Richard, Maarleveld, Joris J., McFadden, Nora, Evans, David J., Simmonds, Peter, 2014. The influence of viral RNA secondary structure on interactions with innate host cell defences. Nucl. Acids Res. 42 (March (5)), 3314–3329.

Züst, Roland, Cervantes-Barragan, Luisa, Habjan, Matthias, Maier, Reinhard, Neuman, Benjamin W., Ziebuhr, John, Szretter, Kristy J., Baker, Susan C., Barchet, Winfried, Diamond, Michael S., Siddell, Stuart G., Ludewig, Burkhard, Thiel, Volker, 2011. Ribose $2'-O-$methylation provides a molecular signature for the distinction of self and non-self mRNA dependent on the RNA sensor Mda5. Nat. Immunol. 12 (February (2)), 137–143.